Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

PERCEPTUAL OPTIMIZATION OF ROOM-IN-ROOM REPRODUCTION WITH SPATIALLY DISTRIBUTED LOUDSPEAKERS

Julian Grosse
Cluster of Excellence "Hearing4all"
Acoustics Group
Carl von Ossietzky University
Oldenburg, Germany
julian.grosse@uni-oldenburg.de

Steven van de Par
Cluster of Excellence "Hearing4all"
Acoustics Group
Carl von Ossietzky University
Oldenburg, Germany
steven.van.de.par@uni-oldenburg.de

ABSTRACT

It is often desirable to reproduce a specific room-acoustic scene, e.g. a concert hall in a playback room, in such a way that the listener has a plausible and authentic spatial impression of the original sound source including the room acoustical properties. In this study a perceptually motivated approach for spatial audio reproduction is developed. This approach optimizes the spatial and monaural cues of the direct and reverberant sound separately. More specifically, the (monaural) spectral cues responsible for the timbre and the (binaural) interaural cross correlation (IACC) cues, responsible for the listener envelopment, were optimized in the playback room to restore the auditory impression of the recording room. The direct sound recorded close to the source is processed with an auditory motivated gammatone filterbank such that the spectral cues, ITD's and ILD's are comparable to the direct sound in the recording room. Additionally, the reverberant sound, which was recorded at two distant locations from the source, is played back via dipole loudspeakers. Due to the arrangement of the two dipole loudspeakers, only the diffuse sound field in the playback room is excited, therefore the spectral cues and the IACC of the reverberant sound field can be controlled independently to match the cues that were present in the recording room. As indicated by a preliminary listening test the applied optimization is perceptually similar to the reference signal and is generally preferred when compared to a conventional room-in-room reproduction.

1. INTRODUCTION

The perception of a sound source strongly depends on the room (e.g. church or a concert hall) in which the source is placed. When a recorded sound source is reproduced, it is desirable to not only faithfully reproduce the sound source but also the room acoustics of the recording room. There are several methods to reproduce a sound field which are based on using large arrays of loudspeakers, e.g. Ambisonics [1] or Wave-field-synthesis (WFS) [2]. Since the above-named methods need a large number of loudspeakers, they are less suitable for sound reproduction in the living room. Furthermore, these approaches assume that the room where the loudspeaker array is placed has no boundaries and the propagating sound wave is not affected by the room. The inaccuracies that will occur due to a reverberative environment are generally not considered.

Some problems that will occur when a sound recorded in a 'recording' room is reproduced in another echoic 'reproduction' room can be understood when considering that in this case that the listener who is present in the reproduction room will effectively hear the combined room acoustics of both rooms. This implies that the Room Impulse Responses (RIR) of both rooms are convolved with one another. As a consequence, the envelope of the resulting impulse response will look like a second order system. Additionally, also the spectral statistics will change. For a single RIR the standard deviation in the magnitude spectrum is approximately $\sigma = 5.5$ dB [3]. Due to the convolution of both Room Impulse Responses the standard deviation of the magnitude spectrum will increase by a factor of $\sqrt{2}$ which may be perceived as an increase in spectral coloration.

This study presents a method that compensates for the detrimental effects of the reproduction room using human auditory perceptual criteria. Thus this method will not attempt to reproduce the sound in an exact physical way at the eardrum of a listener. Instead it optimizes timbre and spatial characteristics based on auditorily motivated frequency bands and on the interaural cross correlation. For optimization an artificial head is placed in the playback room so that the spatial and timbre cues can be matched to a reference artificial head in the recording room. In normal loudspeaker playback, loudspeakers are designed such that the direct sound path has a flat transfer function. As a consequence, there is little control over how the reverberant sound field in the playback room is excited. In our approach, a set of rear dipole loudspeakers will be used the excite the reverberant sound field separately, by aligning the dipole loudspeakers such that the listener receives no direct sound path from the dipole loudspeakers. In this way both the timbre and the spatial properties of the reverberant sound field in the playback room can be controlled separately. This approach has the restriction that the reproduction room needs a smaller reverberation time than the recording room.

Similar perceptually motivated approaches for sound reproduction have been investigated before. De Bruijn et al. [4] showed with a similar setup that it is possible to modify the perceived distance by separately presenting the direct sound over a conventional stereo loudspeaker setup and the reverberant sound over rear loudspeaker. Breebaart et al. [5] found in the context of low-bit-rate audio coding that the binaural attributes ILD (interaural level differences), ITD (interaural time differences) and the IACC (Interaural Cross Correlation) can sufficiently describing the spatial percept of a stereo audio signal. The ILD's and ITD's are responsible for localization and the IACC is responsible for the perceived spaciousness and the listener envelopment [6].

In this study in order to be able to optimize the direct sound and the reverberant sound field separately, the direct sound is recorded close to the sound source in the recording room and is presented

over a stereo loudspeaker setup. The reverberant sound is recorded at two distant positions and is rendered via two dipole loudspeakers. This loudspeaker arrangement gives the possibility to control the overall timbre of the reproduced sound source, the IACC, and the reverberation time of the recording room including the effect of the rendering in the playback room.
In order to evaluate the authenticity of the sound reproduction that can be obtained with the proposed method, the reference signal is compared with the applied optimization over headphones in a MUSHRA-Test (MUltiple Stimulus test with Hidden Reference and Anchor). For additional simulated listening situations, a multichannel reproduction and a more conventional room-in-room reproduction method is used.

2. METHOD

In this section the optimization method is described in detail. An artificial head is placed in the recording room (as the reference) as well as in the playback room (Fig.1) which are used for recording binaural room impulse responses (BRIRs). In this approach, the optimization does not aim for an accurate reproduction of the physical sound field at the ear-drum of the artificial head, but rather the excitation pattern created on the basilar membrane is considered [7]. In this context, the direct and the diffuse sound are optimized separately taking into account the room-acoustical properties of the playback room. In Section 2.1 it will be described how the BRIR is divided into a direct path and a diffuse sound path. In Section 2.2 the analysis of the playback room is described. In Section 2.5 the method is introduced for optimizing the reproduction in the playback room based on perceptual criteria measured at the "ear-drum" of the artificial head.

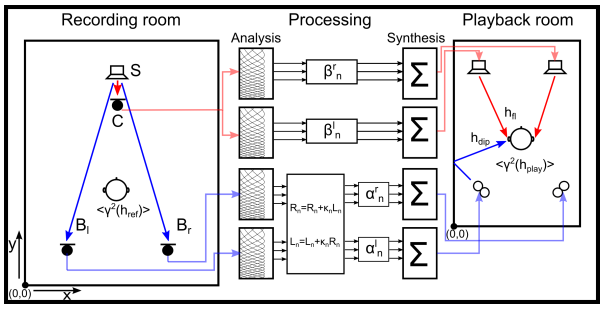

Figure 1: *Left: Recording room with a reference artificial head, the close microphone $C(t)$ for the direct sound and two omnidirectional microphones $B_{l,r}(t)$ in the diffuse sound field. Right: The playback room with the artificial head for the perceptual optimization, two front loudspeakers (60° stereo triangle) and the dipole loudspeakers to excite the diffuse sound field. The processing-stage is divided into an analysis and a synthesis stage. In the analysis-stage the energy outputs of the gammatone filters are observed. Each filter output is multiplied with a real-valued gain factor (β_n for the direct sound $C(t)$, α_n for the reverberant sound $B(t)$) to control the overall coloration. The mixing factor κ_n controls the IACC in the playback room. The synthesis stage includes the phase-alignment of the gammatone filters and finally sums up all sub-bands.*

2.1. Analysis - Recording room

In this section the analysis of two parameters, the timbre and the cross correlation is described. The first parameter is the timbre expressed in terms of the excitation pattern determined for the recording room. The analysis of the the recording room will occur for the direct sound and for the reverberant sound separately to optimize the timbre in the playback room. To clarify the notation conventions, only the left (noted as l) ear of the artificial head is considered in the following. Derivations for the right ear are similar and only require l to be replaced by r and vice versa. If we observe a BRIR in the recording room $h(t)_{ref}$ we can split the BRIR (denoted with the subscript ref) into two parts.

$$h(t)^{(l)}_{ref} = h(t)^{(l)}_{d,ref} + h(t)^{(l)}_{rev,ref} \quad (1)$$

where the indices d, rev are indicating the direct sound and the reverberative sound in the recording room, respectively. The separation of the two parts will occur with the separation time constant t_m. For the separation, a squared cosine window is used with a 4 ms flank. The separation time constant t_m in the optimization will effectively control the direct to diffuse ratio, therefore it controls the T_{60} reverberation time in the playback room. The derivation of the optimal separation time constant can be found in section 3.1. After separating the BRIR, the two parts are filtered with an auditory motivated 4^{th}-order gammatone filterbank (GTFB) (cf. [8]). The filters are distributed equally on an ERB scale (equivalent rectangular bandwidth) in the range of 20 Hz to 24 kHz. This yields 42 gammatone channels. The filtered BRIR signal h in each gammatone channel n is denoted by $\langle\gamma_n(h)\rangle$. The overall excitation pattern is determined by:

$$\langle\gamma^2_n(h(t)^{(l)}_{d,ref})\rangle = \int_{t_d}^{t_m}\int_{-\infty}^{+\infty} |h(\tau)^{(l)}_{ref} * \gamma_n(t-\tau)|^2 d\tau \quad (2)$$

where t_d indicates the start of the impulse response. The excitation pattern of the reverberant part is calculated by integrating the BRIR from t_m until the end of the BRIR. The expression $\langle\gamma^2_n(h(t))\rangle$ contains the energy in each frequency band n at the center frequency f_c of the gammatone filters (called excitation pattern). This excitation pattern control the overall energy in the playback room.
The second parameter analysed is the interaural cross correlation (IACC) in the recording room. The normalized cross correlation coefficient of the whole BRIR is determined by observing the time-signal in each gammatone channel n for the whole BRIR. The IACC is processed in the following way:

$$IACC[q,n] = \frac{\sum_{m=-\infty}^{\infty} h[m,n]_{ref,l} \cdot h[m+q,n]_{ref,r}}{\sqrt{(\sum_{m=-\infty}^{\infty} h[n]^2_{ref,l})} \cdot \sqrt{(\sum_{m=-\infty}^{\infty} h[n]^2_{ref,r})}} \quad (3)$$

in which l and r are the left and right channels of the artificial head and q is the time delay in samples. Within this context the value at $q = 0$ is used.

2.2. Analysis - Playback room

The analysis of the playback room is quite similar to the analysis of the recording room. The complete BRIR in the playback room (denoted with the subscript $play$) is defined by:

$$h(t)^{(l)}_{play} = h(t)^{(l)}_{d,play} + h(t)^{(l)}_{rev,play} \quad (4)$$

The BRIR can again be separated into a direct and a reverberant part. The separation time constant is the same as in Section 2.1. h_{pr} is the BRIR in the playback room (room-in-room (RinR)) when rendering the impulse response measured with microphone C and B. The correction factor β_n is used for the direct sound

in each band and α_n is needed to control the overall energy, thus, the amount of the diffuse field in the playback room. Theoretically, since the recording microphone, $C(t)$, is close to the sound source, only direct sound is recorded, and the RIR is a simple convolution with the BRIR of the loudspeaker to the artificial head in the playback room. We will, however, consider the more general case where $C(t)$ also incorporates some reverberation. In addition the diffuse sound field is excited separately with dipole loudspeakers:

$$h(t)^{(l)}_{pr} = \beta^{(l)}_n [C(t) * h(t)^{(l)}_{play}] + \alpha^{(l)}_n [(B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip})] \quad (5)$$

where the superscript ll refers to the path from the left loudspeaker to the left ear and the superscript rl refers to the path from the right loudspeaker to the left ear of the artificial head. The BRIR $h(t)_{dip}$ of the dipole loudspeakers are convolved with the signals B which were recorded at two distant positions in the recording room. Due to the directivity pattern of a dipole loudspeaker it is possible to excite only the diffuse sound field in the playback room when the zero is directed towards the listener. Because we know that h_{play} can be separated into a direct and a diffuse part we can express Equation 5 as:

$$h(t)^{(l)}_{pr} = \beta^{(l)}_n \left[C(t) * h(t)^{(l)}_{d,play} + C(t) * h(t)^{(l)}_{rev,play}\right] + \alpha^{(l)}_n \left[(B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip})\right] \quad (6)$$

The aim is to make the BRIR measured in the recording and playback rooms equal:

$$h(t)^{(l)}_{ref} = h(t)^{(l)}_{pr} \quad (7)$$

One classical approach would be to equalize the transfer function such that the sound pressure signal at the listeners eardrum is the same in both rooms (like crosstalk cancellation). In the perceptive approach of this study, however, we do not want to optimize the transfer function in an exact physical sense but rather in a physiological sense, i.e. by optimizing the levels measured at the output of the auditory filters. Thus, we only consider the excitation pattern. We can express Equation 7 as:

$$\langle \gamma^2_n (h(t)^{(l)}_{ref}) \rangle = \langle \gamma^2_n (h(t)^{(l)}_{pr}) \rangle \quad (8)$$

Now we can substitute all given impulse responses of the recording and the playback room. By solving Equation 8, the cross terms between the direct sound field and the diffuse sound field are cancelled out because of the assumption that the direct signal are incoherent to the diffuse signal. The resulting final term is shown in Equation 9.

$$\underbrace{\langle \gamma^2_n (h(t)^{(l)}_{d,ref}) \rangle}_{p1} - \beta^2_n \cdot \underbrace{\langle \gamma^2_n (C(t) * (h(t)^{ll}_{d,play} + h(t)^{(rl)}_{d,play})) \rangle}_{p2} + \underbrace{\langle \gamma^2_n (h(t)^{(l)}_{rev,ref}) \rangle}_{p3} - \underbrace{\beta^2_n \cdot \langle \gamma^2_n (C(t) * (h(t)^{(ll)}_{rev,play} + h(t)^{(rl)}_{rev,play})) \rangle}_{p4} - \underbrace{\alpha^2_n \cdot \langle \gamma^2_n ((B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip})) \rangle}_{p5} \stackrel{!}{=} 0 \quad (9)$$

Because the direct sound in the playback room also affects the excited diffuse field in this room (noted in Eq. 9 as $p4$), the direct sound has to be adjusted first. To match the direct sound in the playback room to the reference, the factor β^2_n in Equation 10 has to be processed. The term $p4$ in Eq. 9 where a β^2_n appears does not appear in this equation because the adjusted expression is taken into account when α^2_n is calculated. Equation 10 contains the energy of the direct sound in the recording room and the playback room.

$$\beta^2_{n,l} = \frac{\langle \gamma^2_n (h(t)^{(l)}_{d,ref}) \rangle}{\langle \gamma^2_n (C(t) * (h(t)^{(ll)}_{d,play} + h(t)^{(rl)}_{d,play})) \rangle} \quad (10)$$

The direct sound adjustment makes sure that the energy of direct sound is comparable in both rooms. The overall timbre can now be controlled via the dipole loudspeakers. In Equation 9 the difference between the first two terms $p1$ and $p2$ should be zero because the energy of the direct sound in the playback room was adjusted one step before. This leads in Equation 9 to the following expression for the diffuse sound field:

$$\alpha^2_{n,l} = \frac{\langle \gamma^2_n (h(t)^{(l)}_{rev,ref}) \rangle}{\langle \gamma^2_n (B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip}) \rangle} - \frac{\beta^2_{n,l} \langle \gamma^2_n (C(t) * (h(t)^{(ll)}_{rev,play} + h(t)^{(rl)}_{rev,play})) \rangle}{\langle \gamma^2_n (B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip}) \rangle} \quad (11)$$

An interesting property of Equation 9 is that the term $\langle \gamma^2_n (C(t) * h(t)_{rev,play}) \rangle$ describes the diffuse part of the BRIR in the playback room which is excited by the front loudspeakers. The Equations 10 and 11 can be solved by adapting the excitation pattern of the particular part in the playback room to the excitation pattern in the recording room.

2.3. Coefficient processing

For solving the values of alpha and beta, only the magnitude response is considered. Equation 12 shows exemplarily how the matrix A is computed for the direct path in the playback room.

$$A_{n,f} = |\sum_{n=1}^{P} \gamma_n(f) \cdot H(f)_{d,play}|^2 \quad (12)$$

where each row corresponds to the magnitude transfer function of the gammatone filtered signal. Equation 13 shows the transfer function of the direct path in the recording room.

$$b = |H(f)_{d,ref}|^2 \quad (13)$$

In Equation 14, A is a matrix (known) and α^2 (unknown) and b (known) are vectors.

$$A \cdot \alpha^2 = b \quad (14)$$

If the Matrix A has more rows than columns, the simple solution $\alpha = A^{-1} \cdot b$ can not be applied because A is not a square matrix. In our case we do not have a square Matrix and so we have a overdetermined problem which can be solved using the method of least squares:

$$\alpha^2 = (A^H \cdot A)^{-1} \cdot A^H \cdot b \quad (15)$$

at which superscript H resembles the conjugate transposition of the matrix A. The solution α gives us the gain-factors for each bandpass and can be multiplied in the frequency or time-domain by taking the square root of each element in α^2. In our specific case, the

vector α^2 is the wanted coefficient α_n^2 for the dipole loudspeakers and β_n^2 for the stereo loudspeakers. This solution was suggested by [9].

2.4. IACC optimization

The next step is to optimize the IACC. The correlation strongly depends on the optimization of the direct and diffuse sound. Therefore, the optimization of the IACC is done iteratively by mixing the signals of the omnidirectional microphones in the following way:

$$B_n^{l'} = B_n^l + \kappa_n \cdot B_n^r \tag{16}$$

$$B_n^{r'} = B_n^r + \kappa_n \cdot B_n^l \tag{17}$$

were κ_n is varied iteratively in the range of [-1:1] with a step size of 0.2 in each band n to control the IACC via the dipole loudspeakers. If we apply a $\kappa_n = 1$, the omnidirectional microphone signals B have a maximum correlation. With $\kappa_n = 0$ the signals are mostly decorrelated.

- Step 1: Adjust direct sound such that it is comparable to the direct sound in the recording room (according to Equation 10).
- Step 2: Mix the omnidirectional microphone signals (according to Equation 16 and 17) in the range of [-1:1] in each frequency band n.
- Step 3: Optimize the dipole loudspeaker signals according to Equation 11 that the overall energy in the playback room is comparable to the energy in the recording room.
- Step 4: Comparison of the IACC in the playback room and the IACC in the recording room. (arg $\min(IACC_{rec}(n) - IACC_{play}(n))$)

The iterative process Step 2 to Step 4 is done for every frequency channel. After that, the final processing is made with the best suitable κ_n which minimizes the correlation difference between the recording and the playback room.

2.5. Synthesis

The synthesis stage is used as it was introduced by [8] and is shown in Figure 1. For the synthesis, a 4^{th}-order gammatone filterbank is used with a sampling frequency of 48 kHz. The filters have a bandwidth of 1 ERB (equivalent rectangular bandwidth) between 20 Hz and 20 kHz. This leads to 42 filter coefficients per channel. After processing the coefficients as described in Section 2.2, the coefficients β_n were applied in each gammatone band n as a real-valued gain factor for the direct sound. The same process will occur for the coefficients α_n for the dipole speakers. For the synthesis, the gammatone channels were phase aligned with a delay of 16 ms to avoid audible artefacts in the synthesis stage. The phase alignment is necessary to compensate the physiologically motivated delays of the auditory filters on the basilar membrane. [8] showed that a delay of 16 ms gives good results in this stage. After the phase alignment, the filtered impulse responses are summed across all filter channels P. An example for the direct sound $C(t)$:

$$C(t)_{opt} = \sum_{n=1}^{P} \beta_n \cdot \langle \gamma_n(C(t)) \rangle \tag{18}$$

Now, the direct sound $C(t)_{opt}$ can easily be played back in the playback room via the stereo loudspeakers. For a listening test it is possible to have a headphone reproduction such that a comparison can be made between the reference artificial head signal from the recording room with the artificial head signal of the playback room. This can be achieved by convolving the optimized direct sound with the BRIR of the front loudspeakers. For the headphone reproduction this procedure is done for the dipole loudspeakers, too.

3. RESULTS

3.1. Objective evaluation

In the following section the optimized parameters will be discussed. Figure 2 (top) shows the energy difference of the left artificial head ear between the recording room and the playback room for the simulated lecture room and different conditions. The red curve shows the error between the recording room and the playback room for the perceptual optimization. It can be seen that the fluctuations of the error is fairly small over a wide frequency range. For comparison two conventional recording methods are evaluated also. The first is a Room-in-Room (RinR) rendering which refers to a microphone placed at 2.6 m from the source in the recording room and for which the signal is rendered over two loudspeakers in the playback room. The multi-Channel (mCH) refers to a similar condition where the signals from the omni-directional microphones were rendered on two surround loudspeakers that were placed in the playback room (cf. Fig. 4). The comparison of the rendering methods RinR and mCH with the applied optimization shows that the fluctuations of both methods are greater than the applied optimization. In Figure 2 (middle) the interaural cross correlation is shown. By comparing the IACC of the recording room with the room-in-room (RinR) rendering method it can be seen that the IACC of the room-in-room method is higher over the whole frequency range. The multi-channel reproduction is much lower than the simple RinR-method, but it does not fit to the IACC of the recording room. The comparison of the optimized IACC with the IACC in the recording room shows that both curves fit quite well in most of the gammatone channels. In some channels the IACC cannot reach the desired correlation with simply the dipole signals. For a better adjustment of the correlation, a compensation with the front loudspeaker signals should be taken into account to achieve the reference IACC. In Figure 2 (bottom) the energy decay curve (edc) is illustrated for the recording room and the conditions Opt, RinR and mCH. The RinR-method shows that the edc has a higher descending slope and has a $T_{60,RinR} = 597$ ms. The multi-channel conditions shows a similar slope with a small offset like in the recording room and a $T_{60,mCH} = 695$ ms. The applied optimization shows that the the slope as well as the reverberation time $T_{60,Opt} = 706$ ms is comparable with the reverberation time in the recording room $T_{60,Rec} = 699$ ms.

Figure 3 (top) shows the energy difference of the left artificial head ear between the recording room and the playback room for the simulated church and different conditions. The error of the applied optimization is rather small over a wide frequency range. As can be seen the error between the recording room and the playback room for the conditions RinR and mCh is relatively high, which results from interference of room modes in both rooms. The difference among the RinR and mCh condition is fairly small because of the high front to back ratio of 10 dB, which leads to a similar IACC in Figure 3 (middle). The optimized IACC shows a good agreement with the IACC of the recording room. The energy decay curves illustrate similar properties. The $T_{60,Opt} = 3033$ ms in the playback room are in good agreement with the

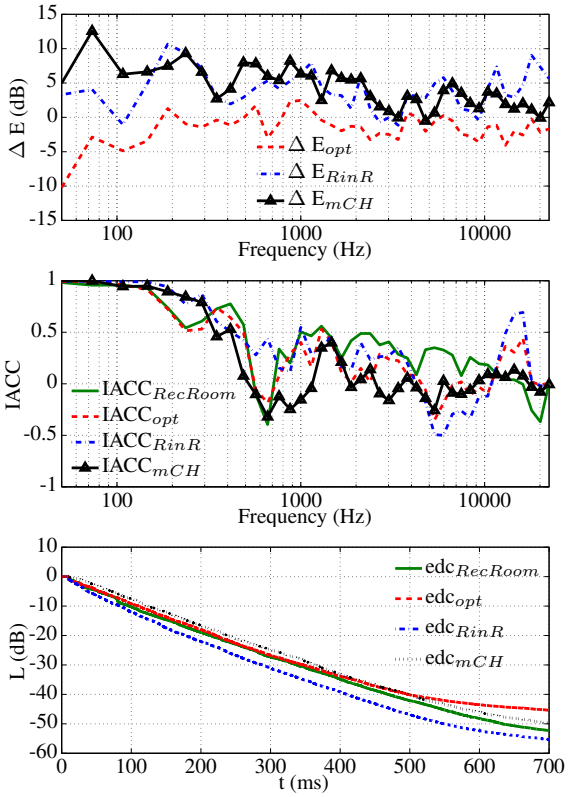

Figure 2: *Comparison of the parameters for the simulated lecture room. Top: Energy difference between the recording room and the playback room for the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi-channel condition (mCh, black). Middle: Illustrated is the IACC of the recording room (green), the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi channel condition (mCh, black). Bottom: Illustrated is the energy decay curve (edc) for the recording room (green), the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi channel condition (mCh, black).*

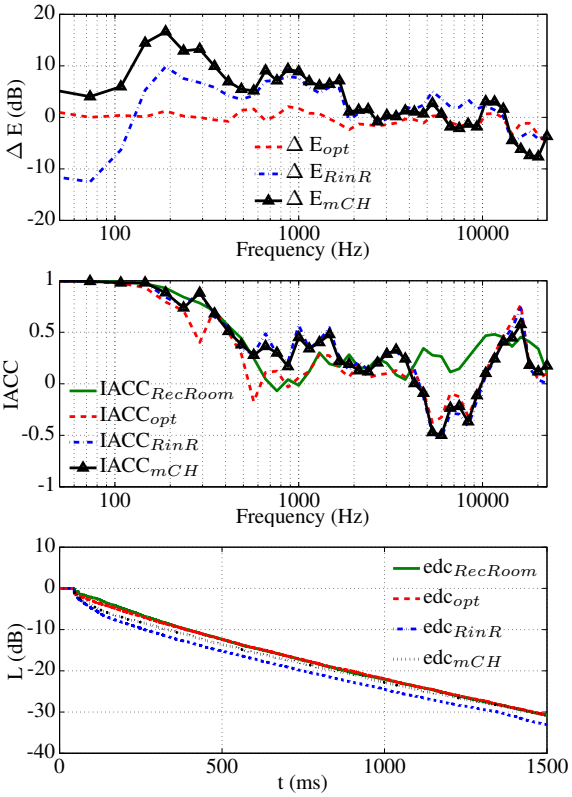

Figure 3: *Comparison of the parameters for the simulated church. Top: Energy difference between the recording room and the playback room for the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi-channel condition (mCh, black). Middle: Illustrated is the IACC of the recording room (green), the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi channel condition (mCh, black). Bottom: Illustrated is the energy decay curve (edc) for the recording room (green), the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi channel condition (mCh, black).*

$T_{60,Rec} = 3040$ ms in the recording room. The multi-channel condition has a $T_{60,mCH} = 2768$ ms and the room-in-room reproduction a $T_{60,RinR} = 2921$ ms. The comparison of the T_{60} reverberation time at the artificial head between the recording room and the playback room with the applied optimization showed that it strongly depends on the separation time constant t_m which was introduced in section 2.1. It was found that for the simulated lecture room, a separation time constant of $t_m = 28$ ms gives a good approximation of the reverberation time of $T_{60,Opt} = 706$ ms which is 7 ms above the reverberation time of the recording room. For the simulated church an separation time constant of $t_m = 60$ ms was found which leads to a reverberation time in the playback room of $T_{60,Opt} = 3033$ ms which is 7 ms below the reverberation time of the recording room. The reproduced T_{60} are below the just noticeable difference of reverberation time, which is in the range of 20% to 30 % [10]. The optimal separation time constant t_m can be derived iteratively by varying t_m from small to bigger values. A small t_m means that only a small amount of the direct sound energy in the recording room is considered which leads to a small amount of direct sound in the playback room. Because the overall energy in the playback room should be comparable to the energy in the recording room, a larger amount of energy has to be rendered over the dipole loudspeaker. This leads to a higher reverberation time in the playback room compared to the T_{60} of the recording room. The optimal constant t_m is derived when the T_{60} in the playback room is comparable to the T_{60} in the recording room.

3.2. Subjective evaluation

In the following section the experimental setup of the listening experiment will be introduced. In this listening test, two recording rooms were simulated. The first room was a lecture room at the University of Oldenburg ($T_{60} = 650$ ms), the second recording room was the St.Marien Church in Oldenburg ($T_{60} = 3040$ ms). The playback room was the loudspeaker lab. ($T_{60} = 400$ ms) at the University of Oldenburg. The loudspeaker orientations used are shown in Figure 4 for the different test conditions. In the subjective evaluation a MUSHRA-Test was used to evaluate the different rendering methods over headphone relative to the reference condition (called ref) in the recording room. In addition, a conven-

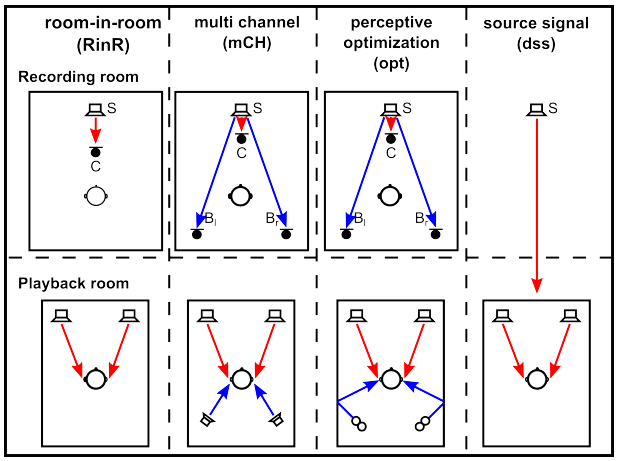

Figure 4: *The four conditions of the listening test. Room-in-room: The microphone signal $C(t)$ (red) is rendered over the front loudspeaker (red) in the playback room. Multi channel: The close microphone signal $C(t)$ (red) is rendered over the front loudspeaker (red) and the microphones $B(t)$ in the diffuse field (blue) are rendered over two rear-loudspeaker (blue). Opt: The optimized close microphone signal $C(t)$ (red) is rendered over the front loudspeaker (red) and the optimized microphones $B(t)$ in the diffuse field (blue) are rendered over the two dipole loudspeaker (blue). dss: This condition contains just the room-acoustical properties of the playback room. The dry source signal is rendered directly in the playback room over the front loudspeaker (red).*

tional room-in-room reproduction (called RinR) method was used. In this condition, the close cardioid microphone C in Figure 4 was recorded in a distance of 2.6 m to simulate a conventional stereo reproduction with a small amount of reverberation. This signal was rendered over the front loudspeaker. For the multi-channel reproduction (mCH) the close cardioid microphone signal C was recorded in a distance of 1.4 m. This signal was rendered over the front loudspeaker like in a 5.1 setup. The signals B_l, B_r was rendered as the two rear-speaker-signals of a 5.1 multi-channel setup. In the multi-channel condition no subwoofer and no center speaker was used. The applied optimization (Opt) was processed like it is described in Section 2 and then rendered in the playback room. As the anchor test-condition (anchor) the reference signal was low-pass filtered at 3.5 kHz as described in [ITU-R BS.1534-1]. The secondary anchor test-condition (called dss) is the dry instrument played back in the playback room. This condition was used to investigate the change in the perceived room-acoustics with respect to the other reproduction methods. To use the source signals of the recording room for the multi-channel reproduction, the energy-ratio of the front channels to the rear channels of four musical DVD's were analyzed. A front to back ratio of 10 dB was found. The close microphone signal of the front loudspeaker was recorded at the same distance which was used in the condition Opt. The dipole-speakers were replaced by two loudspeakers with a conventional directivity pattern (Genelec 6010A) and positioned as described in the [ITU-R BS.775].

3.3. Stimuli and subjects

Twelve different monaural recordings of musical instruments of five to ten seconds in duration were used. The instruments were dry music signals recordings without any room influences. The recordings used were as follows: a piece of Beethoven (recorded by [11]), a choir (recorded by [12]), female speech, a violin (one self recorded and one recorded by [12]), two guitars (chords and picking), clarinet, piano, saxophone, snare drum and a trumpet. All stimuli were presented at 67 dB-SPL. All stimuli were convolved with the room impulse response of the close microphone C and the microphone B in the recording room. These signals were then convolved with the specific binaural loudspeaker impulse response which was measured from the loudspeaker to the artificial head in the playback room (C with the BRIR of the front loudspeakers and B with the BRIR of the dipole loudspeakers). The same procedure was done for the listening conditions RinR and mCh. To have the possibility to compare the recording room with the playback room, the original source signal was convolved with the BRIR of the artificial head of the recording room as a reference signal. The listening test were performed by $N = 12$ normal hearing subjects, nine male and three female, with a mean age of 29 years. Five of twelve participants reported to have musical experience with playing an instrument. The rating was done from all subjects for all conditions and instruments in two sessions. The duration of one session was approximately 60 minutes. The task of the subjects was to rate in a blind test the difference of five processing algorithms (anchor (low-pass filtered at 3.5 kHz), Opt (perceptual optimized room-in-room reproduction), dss (dry signal in playback room), RinR (a conventional room-in-room reproduction) and a multi-channel reproduction) on a scale between 0 (large difference) and 100 (no difference) . Additionally, a hidden reference condition was included. All subjects completed a training phase where all stimulus manipulations were presented for a select number of test stimuli.

3.4. Subjective results

Figure 5 shows the results for the MUSHRA-Test for the lecture room (red) and the church (blue). Illustrated in Figure 5 is the mean over all subjects. The standard error is derived from the mean scores calculated over all subjects and thus shows the variations between the instruments. Examination of the data in Figure 5, it shows that our proposed method, Opt, was always rated with a smaller difference than the conventional room-in-room reproduction (RinR). This trend can be seen for the lecture room as well as for the church. The perceived difference could be caused by the stronger variations in energy (which cause an increase in coloration), a much higher IACC over all frequencies and a lower reverberation time which are illustrated in Figure 2. A comparison of the results of the condition Opt with the multi-channel condition (mCH) shows that for the lecture room the perceived difference is comparable. This can be seen in Figure 2, that the Energy Decay Curve of this condition is comparable to EDC of the recording room as well as the IACC. Comparing the conditions Opt and mCh of the church, it shows that the condition Opt was rated with no difference. The condition mCh shows that it was rated much lower as the RinR condition in the church. The reason why the condition RinR is rated much higher than the multi-channel (mCh) reproduction could be explained considering that the front to back ratio of 10 dB is too high. A reason for this is that the distance of the close microphone signal is closer at the sound source than the RinR condition. Therefore, the mCH condition has less reverberation in the close microphone signal, which cannot be compensated by the rear-speaker in the mCH condition. In the multi-channel condition, the ratio between the front and back channel signals control the direct to diffuse ratio of the rendered signal. The reproduced

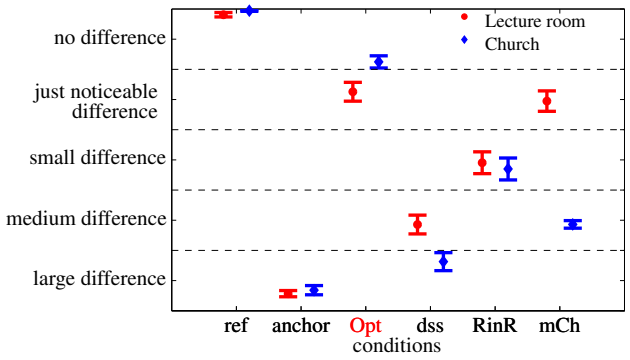

Figure 5: *Subjective measurement results for the headphone reproduction of the lecture room (red dots) and the church (blue diamonds) in a MUSHRA-Test. The symbols are the mean over 12 instruments and 12 subjects. The error bars are the standard error and indicates the variation over the 12 instruments. The x-axis are the different processing conditions, the y-axis indicates the difference on a scale between* 0 *(large difference) and* 100 *(no difference). Our proposed method (Opt) is marked in red on the x-axis.*

signals in the condition (mCH) are less reverberant than the condition RinR. The condition dss shows the dry source signal which was played directly in the playback room. This condition shows how large the perceived difference of the recording room is compared to the playback room. This condition should be rated much lower than the other conditions (with the exception of the 3.5 kHz anchor signal (anchor)) because only the room-acoustical properties of the playback room is included. In addition to the previous listening test, a live listening test over loudspeaker should be performed, to validate the previous results. This could be necessary to include the effects of head movements and the individual head-related-transfer-functions of the listener. Olive et al. ([13]) compared a live loudspeaker reproduction with a binaural reproduction over headphones in a subjective listening test. They showed that the scores between a live representation and a headphone representation have minor discrepancies which could result out of the removal of the visual biases and head movements. However, it can be seen that the standard errors are in a fairly small range and our proposed method works quite well over all stimuli used.

4. CONCLUSIONS

In this study a method was presented for rendering the room acoustics of a recording room in an echoic playback room. This method compensates for the reproduction conditions in the playback room. Rather than attempting to recreate the physical sound field, the proposed method optimizes the perceptual attributes IACC and the overall timbre in a playback room using a fairly small amount of loudspeakers. Because of the placement of an artificial head at the recording side it is possible to analyse the specific room dependent timbre and binaural cues and reproduce these on the playback side. For sound reproduction in the playback room, a conventional stereo loudspeaker setup is first used to reproduce the direct sound. With this setup, we can control the direction of arrival and the amount of energy in the playback room corresponding to the direct energy in the recording room. Furthermore the energy difference as well as the IACC can controlled with diffuse sound via a set of dipole loudspeakers. Because of the directivity pattern only the diffuse field is excited which implies that the dipoles are not perceived as a separate sound source provided that the head is in the sweet spot. For a better understanding in terms of head movements with this setup, a new listening test which covers various listening positions should be conducted. A comparison of the subjective ratings in Section 3.4 showed a higher preference for our proposed method with reference to the conventional room-in-room reproduction.

5. ACKNOWLEDGMENTS

This work was supported by the DFG Forschergruppe Individualisierte Hoerakustik (FOR-1732). The author thanks the two anonymous reviewers for the comment, that helped to improve the paper.

6. REFERENCES

[1] M.A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10, 1973.

[2] A.J. Berkhout, "A holographic approach to acoustic control," *J. Audio Eng. Soc.*, vol. 36, pp. 977–995, 1988.

[3] M.R. Schroeder, "Statistical parameters of the frequency response curves of large room," *J. Audio Eng. Soc.*, vol. 35, pp. 299–306, 1987.

[4] W. de Bruijn, A. Härmä, and S. van de Par, "On the use of directional loudspeakers to create a sound source close to the listener," in *124th AES Convention*, May 2008, Amsterdam, Netherlands.

[5] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *Journal on Applied Signal Processing*, vol. 9, pp. 1305–1322, 2005.

[6] J.S. Bradley and G.A. Soulodre, "Objective measures of listener envelopment," *J. Acoust. Soc. Am.*, vol. 98, pp. 2590–2597, 1995.

[7] Brian C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th edition edition, 2 January 2012.

[8] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," Tech. Rep., Acta Acustica united with Acustica , Vol. 88, pp. 433–442, 2002.

[9] J. Dattorro, "Constrained least squares fit of a filter bank to an arbitrary magnitude frequency response," 1991.

[10] Zihou Meng, Fengjie Zhao, and Mu He, "The just noticeable difference of noise length and reverberation perception," in *Communications and Information Technologies, ISCIT '06. International Symposium on,*, 2006, pp. 418 – 421.

[11] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, pp. 856–865, 2008.

[12] R. Freiheit, "Creating an anechoic choral recording," in *Proc. of the International Symposium on Room Acoustics*, 2010.

[13] Sean E. Olive and Peter L. Schuck, "The effects of loudspeaker placement on listener preference ratings," *J. Audio Eng. Soc.*, vol. 42, no. 9, pp. 651 – 669, 1994.

[14] O. Warusfel and N. Misdariis, "Directivity synthesis with a 3d array of loudspeakers application for stage performance," *Proceedings of the COST G-6 Conference on Digital Audio Effects*, vol. DAFX-01, pp. 1–5, 2001.